Extending and Testing the Bayesian Theory of Generalization

Daniel J. Navarro (daniel.navarro@adelaide.edu.au)
School of Psychology, University of Adelaide, SA 5005, Australia

Michael D. Lee (mdlee@uci.edu)
Department of Cognitive Science, UC Irvine, CA 92697, USA

Matthew J. Dry (matt.dry@psy.kuleuven.be)
Department of Psychology, Leuven University, Tiensestraat 102, B-3000 Leuven, Belgium

Benjamin Schultz (benjamin.schultz@adelaide.edu.au)
School of Psychology, University of Adelaide, SA 5005, Australia

Abstract

We introduce a tractable family of Bayesian generalization functions. The family extends the basic model proposed by Tenenbaum and Griffiths (2001), allowing richer variation in sampling assumptions and prior beliefs. We derive analytic expressions for these generalization functions, and provide an explicit model for experimental data. We then present an experiment that tests the basic model predictions within the core domain of the theory, namely tasks that require people to make inductive judgments about whether some property holds for novel items. Analysis of the results illustrates the importance of describing variations in people's prior beliefs and assumptions about how items are sampled and of having an explicit model for the entire task.

Keywords: generalization, induction, Bayesian models

The ability to recognize that a novel item shares unobserved characteristics with items previously encountered is an extremely useful inductive capacity, so it is not surprising that psychologists devoted some effort to understanding how people make these generalizations. The most well-known account of simple generalizations is Shepard's (1987) exponential law, derived from a Bayesian analysis and experimental work dating back to the 1950s (e.g., Shepard, 1957). According to Shepard's analysis, the learner assumes that there exists some unknown *consequential region* within an appropriate psychological space, and that generalization probabilities result from the learner integrating over his or her uncertainty about the boundaries of the region.

Although a considerable body of modelling work relies on Shepard's law to justify the use of exponential functions, few researchers have sought to apply or extend his analysis on its own terms (see, e.g., Navarro, 2006, for a discussion). The major exception to this is Tenenbaum and Griffiths (2001), who introduce three innovations: firstly, they note that the basic Bayesian machinery can easily handle multiple training examples, and that it is merely analytic intractability that has prevented people from doing so previously. Secondly, they note that the approach can be extended to non-spatial representations, and in doing so make connections to Tversky's (1977) featural approach. Thirdly, they note that variation in prior beliefs, assumptions about how stimuli are sampled, and the nature of the hypotheses involved (e.g., connected versus disconnected regions; Navarro, 2006) induce a number of interesting changes to the model. Remarkably, however, there are few formal results or experimental data that allow these extensions to be explored as well as one might like: consequently, our goal in this paper is to provide both. In the first part of the paper, we present the theoretical extensions, while the second half is devoted to experimental data and its analysis using these extensions.

Modelling Generalization

This section extends the Bayesian theory of generalization in four ways, by (1) expanding the range of allowable sampling assumptions, (2) explicitly allowing variability in prior beliefs, (3) providing analytic expressions for the resulting generalization gradients, and (4) including task-specific statistical models for calibration, contaminants and errors.

Sampling Assumptions

Bayesian generalization models assume that if some property holds for previously observed items $X = (x_1, \ldots, x_n)$ then they may all be taken to belong to some latent region r over which the property holds. As a result, the inductive problem when presented with new item y is to infer $p(y \in r|X, X \in r)$, the probability that the new item also belongs to the region. Note that this induction uses two pieces of knowledge: that (a) the items belong to the region (i.e. $X \in r$), and (b) the items have really been observed (i.e., X exists). The first fact implies an obvious constraint on the region boundaries, but the second is more subtle. In Shepard's original proposal (weak sampling), the generative process is assumed to be independent of r, so the probability of sampling an item x such that $x \in r$ is a constant, $p(x, x \in r|r) \propto p(x) \propto 1$ if $x \in r$, and 0 otherwise. In contrast, Tenenbaum and Griffiths' strong sampling proposal states that the observations are explicitly sampled from the region (with uniform probability density on r), implying that $p(x, x \in r|r) = 1/|r|$ if $x \in r$ where $|r|$ denotes the size of the region, and is 0 otherwise.

In our view, strong and weak sampling are best viewed as two end points on a continuum: at one end the training items are sampled in a way that is completely dependent on the region itself, whereas at the other end observations are completely independent of the consequence at hand. However, in many realistic scenarios our observations arrive in a manner that is only partially correlated with the phenomenon in which we are interested. As a simple example, consider the sampling process involved when one is trying to guess whether a patient in a doctor's office is sick. Not everyone who enters the office is in fact sick, so strong sampling is impossible. However, people who are seeking treatment *are* more likely to be sick than randomly chosen people, so weak sampling seems inappropriate too. In short, a more general approach is necessary. Perhaps the simplest scheme that satisfies this

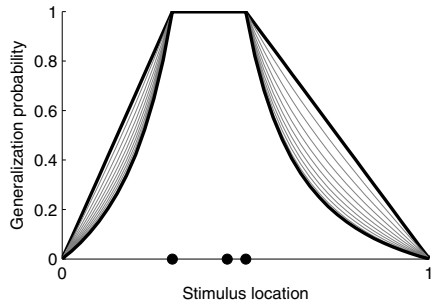

Figure 1: The effect of varying θ, for a case involving three training items (black dots) varying along a single dimension. In this example the region is known not to extend below 0 or above 1, but in all other respects the prior over regions is uniform. When $\theta = 0$ (weak sampling), we obtain a linear interpolation model (the uppermost black curve), whereas when $\theta = 1$ (strong sampling), we obtain the tightest generalization gradients (the lowest black curve). Varying θ in increments of 0.1 produces the various intermediate gradients shown with the grey curves.

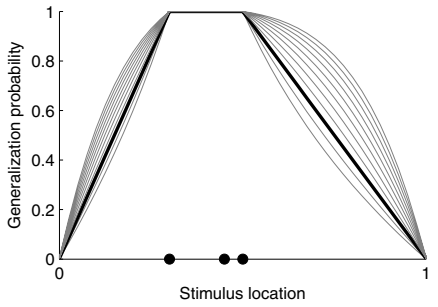

Figure 2: The effect of varying ϕ, when the sampling model is weak $\theta = 0$ for a case involving three training items (black dots). When $\phi = 1$ (the black curve), the linear interpolation function for weak sampling is obtained. When $\phi < 1$, the gradients become convex and dip below the linear one, whereas when $\phi > 1$ the gradients become concave. An analogous effect exists for all values of θ.

criterion is a *mixed sampling* approach; with probability θ, items are sampled from the region in question, but with probability $(1 - \theta)$ they are generated randomly. In reality, this is probably still too simple (in a doctor's surgery, for instance, the non-sick people do tend to *look* sick), but it is nevertheless considerably more useful than the simple strong-versus-weak dichotomy. This generalized sampling model assumes that the probability of sampling item x such that $x \in r$ is

$$p(x, x \in r|r, \theta) = (1 - \theta) + \theta|r|^{-1}. \quad (1)$$

Not surprisingly, the generalization functions that arise from this class of sampling assumptions interpolate smoothly from weak to strong sampling, as shown in Figure 1.

Prior Beliefs

In order to produce generalization gradients shown in Figure 1, we made two additional assumptions. Firstly, we assumed that the region does not extend beyond a finite range (helpful for both experimental and analytic purposes), which without loss of generality we fix at $[0, 1]$. Secondly, we assumed that so long as this constraint is met, every region is equally plausible a priori. It is this latter, rather unrealistic restriction that we now relax, and introduce a simple class of priors indexed by a single parameter ϕ. This prior is intuitively reasonable to the extent that it allows preference for large regions ($\phi > 1$), small regions ($\phi < 1$), or no preference at all ($\phi = 1$), but is nevertheless more restricted than what people's "real" beliefs might encompass, since does not allow a prior preference for "medium sized" regions, or any more complicated beliefs (e.g., "big or small, but not medium").

To construct this family, we first make the assumption that the prior over regions is *location invariant*.[1] In a single dimension, a region is defined in terms of an upper bound u and lower bound l, but may also be described in terms of the mean $(u+l)/2$ and the size $u-l$. Location invariance implies that $p(r) \propto p(u - l)$. For a generalization within the interval $[0, 1]$ we adopt the one-parameter Beta$(1, \phi)$ family, in which $p(u - l) \propto (u - l)^{\phi-1}$, for $u - l \in [0, 1]$. Importantly, since $|r| = u - l$, this prior has the same structure as the likelihood, allowing ϕ to be easily interpreted (e.g., increasing ϕ by one has a similar effect to decreasing the sample size by one).

The effect of allowing a range of priors is illustrated in Figure 2, in which ϕ varies from 0 to 5 in increments of 0.5. As noted, varying the prior has an effect not dissimilar to varying the likelihood function or adding data. Moreover, it is important to recognize that when there is a prior expectation that the region will be large (i.e., $\phi > 1$), the generalization gradients can in fact be *concave*. Accordingly, careful experimental design is required to discriminate between the effects of the two parameters: specifically, a *single* generalization gradient cannot differentiate between ϕ and θ. To disentangle priors from likelihoods, one needs to examine how generalization functions change as new observations are added.

Mathematical Details

In this section we briefly demonstrate the manner in which analytic expressions may be obtained for the generalization function. Space constraints require us to present only a sketch of the derivation, but the full version is available in an accompanying technical note (Navarro, 2008). We are interested in the case where items vary continuously along the finite range $[0, 1]$, and the learner applies the extended generalization model introduced above. Having observed a set of items $X = (x_1, \ldots, x_n)$ such that all items fall inside an unknown region $X \in r$, the learner observes that the probability that a novel item y also falls inside r is

$$\begin{aligned} &p(y \in r|X, X \in r) \\ &\quad = \textstyle\int_{\mathcal{R}} p(y \in r')p(r = r'|X, X \in r)\, dr'. \end{aligned} \quad (2)$$

In this expression, r' denotes one possibility as to the identity of the unknown region r, and the integration is taken over $\mathcal{R}$, the set of all such regions. Noting that $p(y \in r')$ is a simple indicator function that equals 1 if y falls inside r' and 0 if it does not, the application of Bayes' rule yields the expression:

$$p(y \in r|X, X \in r) = \frac{\int_{\mathcal{R}_y} p(X, X \in r|r')p(r')\, dr'}{\int_{\mathcal{R}} p(X, X \in r|r')p(r')\, dr'}. \quad (3)$$

[1] Note that a uniform prior over region locations *does not* imply symmetric or location-invariant generalization gradients, since a shift in location alters the information provided by the edge points.

Letting $z_l = \min(x_1, \ldots, x_n)$ and $z_u = \max(x_1, \ldots, x_n)$ denote the most extreme of the observed data points, it is clear that the denominator of Eq. 3 integrates over regions with upper bound u and lower bound l such that $l \leq z_l$ and $z_u \leq u$. The numerator is more stringent, requiring also that y fall inside the region, so the domain consists of regions satisfying $l \leq \min(z_l, y)$ and $\max(z_u, y) \leq u$.

Under weak sampling, the likelihood function is a simple indicator function that assigns constant probability to any observations X that fall within region r'. With strong sampling, the probability is scaled by size, with the likelihood being given by $|r'|^{-1} = (u-l)^{-1}$ for each observation. With mixed sampling, either of these two possibilities could hold for any particular data point. As a result, the number of "strongly" sampled items in the training set follows a Binomial(θ, n) distribution, which gives rise to the more general likelihood:

$$p(X, X \in r|r, \theta) = \sum_{k=0}^{n} \binom{n}{k} (1-\theta)^k \theta^{n-k} (u-l)^{-(n-k)}. \quad (4)$$

We then substitute this likelihood function and the prior $p(r) \propto (u-l)^{\phi-1}$ into Eq. 3. By cancelling constant terms and rearranging, it is easy to show that the exact two-parameter generalization function is

$$p(y \in r|X, X \in r, \theta, \phi) = \frac{\sum_{k=0}^{n} b(n,k,\theta) f(n-k-\phi+1, \min(y, z_l), \max(y, z_u))}{\sum_{k=0}^{n} b(n,k,\theta) f(n-k-\phi+1, z_l, z_u)}, \quad (5)$$

where $b(n, k, \theta) = \binom{n}{k} (1-\theta)^k \theta^{n-k}$ is the probability that exactly k of the n observed items were sampled weakly, and

$$f(w, a, b) = \int_0^a \int_b^1 (u-l)^{-w} \, du \, dl, \quad (6)$$

for $0 \leq a \leq b \leq 1$. Since the integrand in Eq. 6 is polynomial in u and l, it is trivial to solve analytically, but the expressions are lengthy (see Navarro 2008). In short, the generalization probabilities may be computed exactly as the ratio of the two sums in Eq. 5, though for large n further simplifications (e.g., Gaussian approximation to the binomial) may be useful.

Completing the Model

To complete the model, we need to address several topics that, though somewhat ancillary to the underlying theory of generalization, are essential for the proper representation of experimental data. With this in mind, we briefly outline our approach to (a) response biases, (b) errors and outliers, (c) model evaluation, and (d) individual differences.

Biased probability judgments. The first issue to note is that the generalization function describes a *latent* subjective probability, and people may not always report this value in a straightforward fashion. In the context of the Bayesian model, it is helpful to note that the probability in question is essentially a subjective *confidence* that some rule holds. With this in mind, it makes sense to assume that the function relating the true probability $p(y \in r|X, X \in r)$ to the value $\tilde{p}_y$ that one might expect to see reported is much the same as a "confidence calibration" curve measured in the decision-making literature, which appear to be approximately linear (e.g., Baranski & Petrusic, 1998, Weber & Brewer, 2004). Accordingly, while we might hope calibration to be fairly good in simple inductive tasks, it would be sensible to adopt the assumption that

$$\tilde{p}_y = j_l + (j_u - j_l) p(y \in r|X, X \in r) \quad (7)$$

where the function is parameterized by unknown calibration parameters j_u and j_l, the upper and lower bounds on values that the participant is willing to report when making probability judgments.

Errors and outliers. Once the linear calibration function is incorporated, we have a reasonably plausible model for the most likely response. Nevertheless, since data are noisy, an explicit error model is required. Note that since responses vary continuously between 0 and 1, the standard homoscedastic Gaussian model (used when minimizing squared error) is inappropriate, since the boundaries introduce skewed errors. A more plausible approach is to assume that errors are Beta-distributed such that the most likely response is $\tilde{p}_y$. Accordingly, we specify a skewed error model using the Beta$(1+\tilde{p}\tau, 1+(1-\tilde{p})\tau)$ distribution[2], in which τ is a precision parameter, and the distribution becomes more skewed as $\tilde{p}$ moves toward either 0 or 1. However, this error model does not account for genuine *contaminant processes*: sometimes people give arbitrary responses due to inattention, accidental responding, or any of a range of possibilities. The result is that in such cases the response is entirely independent of the model, and very likely to produce genuine outlier data that can distort the parameter estimates. Accordingly, we assume that with some unknown (but presumably small) probability ϵ, the response is sampled from a uniform distribution on $[0, 1]$.

Model evaluation. Our overall approach to model evaluation is pragmatic. We adopt a simple Bayesian approach for parameter estimation, setting priors over the parameters and then selecting the posterior mode as the best-fitting parameter set. For model checking, however, we rely on orthodox methods (primarily Kolmogorov-Smirnov tests of distributional equivalence) to ascertain whether the posterior mode provides a sufficiently good account of the data. The approach to choosing priors uses a mix of objectivist and subjectivist Bayesian methods, though space constraints preclude a detailed exposition (see Navarro 2008). Briefly, we adopt a uniform prior over the sampling models, $p(\theta) = 1$ and set $p(\phi) \propto \phi \exp(-\phi)$ to ensure that the prior mode involves no preference for region size (i.e., $\phi = 1$). The prior on the calibration function has a weak bias towards an assumption of perfect calibration, in which $p(j_u, j_l) \propto j_u(1-j_l)$ subject to the constraint that $j_l < j_u$ to set the prior over the calibration function. The prior on the precision $p(\tau) \propto (\tau+3)^{-3/2}$ is chosen so as to assume an approximately uniform prior over the standard deviation of the errors. Finally, we use a very tight prior $p(\epsilon) \propto \epsilon^{-1/2}$ over the contaminant probability so as not to encourage the model to "throw away" too many observations as outliers.

[2]That said, though it seems to improve on the Gaussian model, the Beta still has problems very near to the edge points, so in practice the data are truncated to fall on $[.01, .99]$.

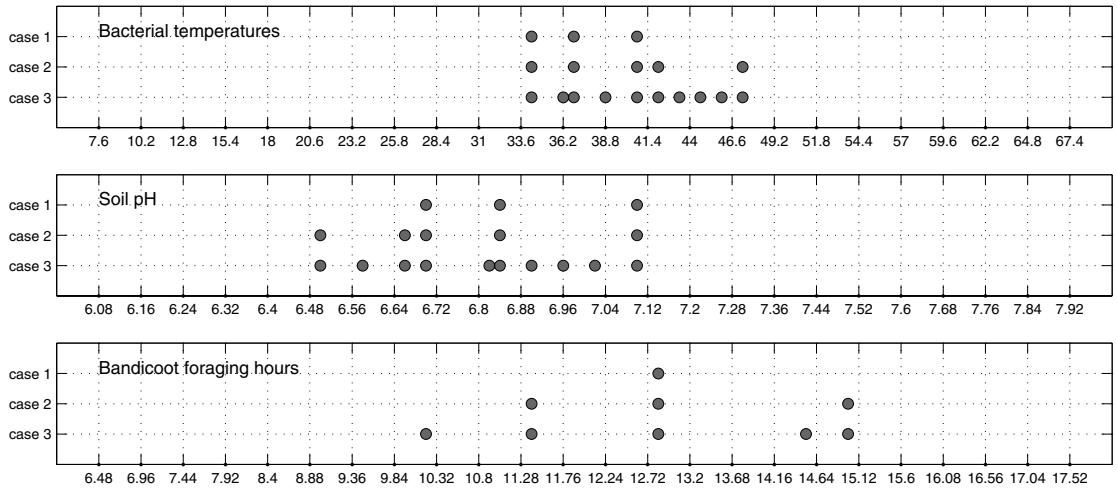

Figure 3: The experimental design. Each panel corresponds to one of the three scenarios, and shows the three different sets of stimuli known to possess the property (circles, squares and triangles). The tick marks are located at each of the test points.

Individual differences. In general, we have no strong reason to assume homogeneity among participants or across different situations, but we do wish to assume that parameter values do not vary with sample size. Although more sophisticated methods are possible (e.g., Navarro, Griffiths, Steyvers, & Lee, 2006), for the present purposes we estimate a separate set of parameters $(\theta, \phi, j_l, j_u, \tau, \epsilon)$ for each scenario and each person, but require the parameters to remain invariant as the number of observations changes.

Qualitative remarks on model complexity. Given the natural concerns one might have regarding model complexity (e.g., Myung, 2000), it is worth commenting briefly on what characteristics the model can and cannot produce. In particular, the following qualitative constraints appear to be the most important: gradients must be unimodal, may not become shallower as more observations arrive, and must remain flat across the region spanned by the observations.

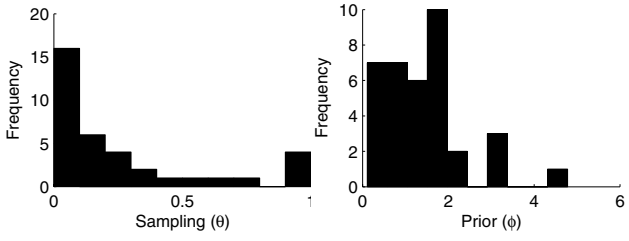

Figure 4: Marginal distributions over θ and ϕ for the 36 model-consistent cases. The two distributions are weakly correlated.

Experiment

Method

Twenty-two undergraduate participants (16 female, 6 male) were asked to evaluate three different generalization scenarios and given a $10 book voucher for their participation. The three scenarios involved different problems in a biological domain: in one case the problem involved the temperatures at which bacterium can survive, in another the range of soil acidity that produces a particular colored flower, and the third related to the times at which a nocturnal animal might forage. Observations were presented on a computer screen as black dots, and participants were asked to solve an induction problem such as the following:

> *Bacillus cereus* is a type of bacteria that causes food-poisoning. This bacteria is sensitive to temperature, and exposure to very high temperatures (>70°C) or very low temperatures (<5°C) will quickly kill the bacteria. In an experiment, food was contaminated with *Bacillus cereus* and then either heated or chilled to a given temperature.
>
> If the experiment found that the bacteria was alive in food that was kept at the temperatures shown as black dots below, what is the probability that it would also be found alive in food kept at the temperature specified by the red question mark?

Responses were obtained by allowing people to position a slider bar using the mouse. In this bacterial scenario, three known observations were initially given, and to elicit the full generalization gradients the question was repeated 24 times, on each occasion asking about a different temperature (in a randomized order). Once this was complete, two new data points were added and the process repeated. Finally, a further five data points were added, and a third generalization gradient elicited. This sequence is illustrated in the top panel of Figure 3. A similar process applied to the soil and foraging scenarios, with the locations of the training points shown in the lower two panels of Figure 3. Note that the relative positions of the test points (i.e., the red question marks) were the same in every single case, though the presentation order differed each time. The three scenarios varied slightly in terms of the extent to which the edge-points were made explicit (e.g., the temperature range explicitly states 5-70°C, whereas the foraging scenario marked 6am-6pm on screen, but only explicitly referred to "night time" as the relevant range) and were presented in a random order.

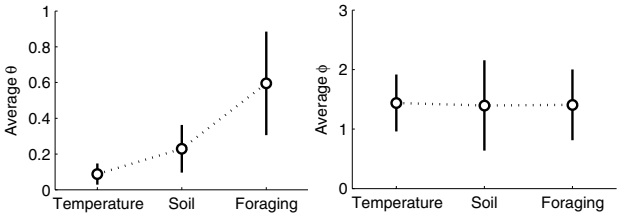

Figure 5: Average θ and ϕ across scenarios, for the 36 model-consistent cases. Error bars are 95% confidence intervals.

Results

Calibration and error. With 22 participants and 3 scenarios, 66 independent parameter optimizations were required, each requiring 6 unknowns to be estimated from 72 data points. Although the parameters of interest are θ and ϕ, we begin with the various precision (τ) and contamination (ϵ) parameters. Overall, the data appear largely uncontaminated, with 45 of the 66 ϵ values less than .001. The precision was reasonable, with the distribution over τ such that the average standard deviation of the error distributions was 0.11. As expected, calibration was generally good but not perfect. At the top of the scale, only 5 cases involved $j_u < .9$. At the lower end some miscalibration is evident, with 24 cases involving $j_l > .1$ (interquartile range ran from .01 to .28). However, much of this variation may represent model misspecification, in the sense that if people do not believe that generalization probability is zero at the edge points, wider gradients are observed; an effect that can be mimicked by raising j_l.

Model checking. Before proceeding to a discussion of the estimates of θ and ϕ, it is important to check that the model provides a good enough description of the data that these estimates are likely to be useful. Given that the model is such that the error distribution is different for every data point, this is not entirely simple. However, since it is straightforward to compute the inverse cumulative distribution functions for the Beta-error model, we can obtain the theoretical percentile rank for each datum. If the model performance is accurate, these should be uniformly distributed, which may be checked via the Kolmogorov-Smirnov test. We conducted these tests at three different levels of granularity. At the lowest level, we checked each of the $22\times3\times3 = 198$ gradients separately: at a significance level of $\alpha = .05$, 66 of the 198 theoretical gradients were rejected. A stricter test would treat each of the 66 parameter estimates separately, and require all three generalization gradients to pass (at the adjusted level α_3, where $1 - \alpha_k = (1 - \alpha)^k$ to hold the error rate fixed at α). This analysis suggests that 30 of the parameter estimates may be unreliable, since at least one generalization gradient was not successfully accounted for. At the most stringent level, we treated each participant separately, requiring all 9 generalization gradients to produced by the participant to be correctly described. In this case, the model passed only for 7 of the 22 participants: however, since "failure" here refers to the inability to fully describe the joint distribution over 216 dependent variables (one per response), a 32% success rate is actually quite good.

Samplers, priors and stories. Restricting the discussion to those 36 parameter estimates that pass the Kolmogorov-Smirnov test (i.e., the $66 - 30 = 36$ cases not rejected), Figures 4 and 5 provide an illustration of the basic pattern of variation. As shown in Figure 4, the estimates of θ tended to be low (mean = .25, std = .33), suggesting a fairly weak degree of correlation between the sampling process and the underlying region. However, the distribution is somewhat bimodal, with a small peak at $\theta = 1$. For ϕ, the distribution is unimodal and slightly skewed (skew = 1.24), with moderate variance (std = 1.03). With a median at $\phi = 1.17$ (mean = 1.42), the general tendency is towards flat priors, but with enough variation to matter for small samples (recall that altering ϕ by 1 has a similar effect to raising the sample size by 1). Moreover, the two distributions are weakly correlated, with $\rho = .33$ ($p \approx 2.8 \times 10^{-5}$): stronger sampling is weakly associated with prior preference to larger regions. Finally, as shown in Figure 5, the different scenarios did appear to suggest different sampling models to people (t-tests are significant at $p < .05$ in all cases), but did not influence the prior beliefs about regions ($p > .9$ in all cases).

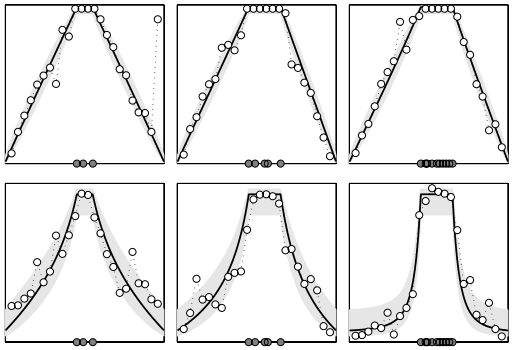

Figure 6: A comparison between partipants 13 (top) and 15 (bottom) on scenario 1 (temperature), as the sample size is increased (from left to right). The solid black lines are the theoretical generalization gradients, with 80% confidence bands for a single judgment shown in grey. White circles denote the actual responses given, and the black circles show the observations given to the participants. Note that both participants are well calibrated, with lower bounds $j_{l,13} = .01$ and $j_{l,15} = .07$ and upper bounds $j_{u,13} = .99$ and $j_{u,15} = .93$, and use a prior slightly favouring large regions ($\phi_{13} = 1.15$, $\phi_{15} = 1.71$). Participant 13 applies a weak sampling model $\theta_{13} = .016$ while participant 15 adopts an intermediate approach $\theta_{15} = .47$. The comparison also highlights the different roles played by τ and ϵ: data in the top panels are precise $\tau = 68$ with a few contaminants $\epsilon = .05$, whereas the data below are uncontaminated $\epsilon = 0$ but less precise $\tau = 19$.

Discussion

The analyses presented make clear that people differ in their assumptions about how observations are generated, and have different prior beliefs that influence the generalization function; but that, nevertheless, the extended Bayesian approach performs well. To make this more concrete, it may help to consider two specific comparisons. The first, illustrated in Figure 6, involves participants 13 and 15 and the temperature scenario, who differ primarily in terms of the sampling assumptions (see Figure caption for the specifics). When only three data points are available (left panels), the two produce very similar gradients. For participant 13, for whom $\theta = .016$, the generalization gradients do not narrow as the

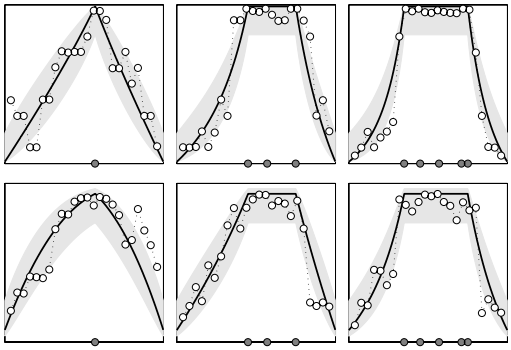

Figure 7: Comparing participants 10 (top) and 4 (bottom) on the foraging scenario. This scenario tended to produce high θ values, with $\theta = 1$ for both participants shown. Both participants are well-calibrated, with $[j_l, j_u]$ equal to $[.01, .99]$ and $[.07, .93]$ respectively. Both are moderately clean data sets, with τ equal to 10.9 and 11.8 respectively, and $\epsilon = 0$ in both cases. However, while $\phi = 1.8$ for participant 10, $\phi = 3.3$ for participant 4.

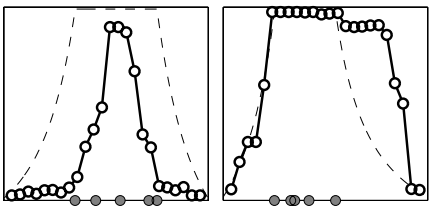

Figure 8: Two cases that cannot be captured by the model. Model predictions are shown as dashed lines, while the solid lines show the data. The left panel corresponds to participant 15, foraging scenario, case 3; and the right panel is participant 10, soil story, case 2.

sample size increases (from left to right). However, since participant 15 assumes correlated sampling (with $\theta = .47$), the gradients tighten considerably from left to right.

The second comparison involves participants 10 and 4, who both apply strong sampling assumptions ($\theta = 1$) to the foraging scenario, as shown in Figure 7. As before, the data are fairly precise and comparable in terms of calibration (see Figure caption). In this case, the main difference lies in the prior: participant 10 has a slight prior bias to favor large regions ($\phi = 1.8$), whereas participant 4 has a large bias ($\phi = 3.3$). As a consequence, the gradients in the lower panel start concave, and only assume a convex shape after multiple data are observed. Note also that, as is generally the case with Bayesian models, the data "swamp" the prior. In Figure 6, participants made different assumptions about sampling, and so grew more dissimilar as the sample size increased. However, in this second comparison, participants agree about how data are produced: as a consequence, their prior differences are erased as the sample size increases from left to right.

Finally, it would be remiss not to discuss the characteristics of those generalization functions that are *not* well-captured by the Bayesian model. In some cases, the reason for this is simply that the data are too noisy to model effectively. In others, the individual generalization curves are consistent with the model, but the variation across cases (i.e., as data are added) are not consistent with this particular model. However, in some cases, there is clear qualitative evidence that participants gave sensible answers that are simply outside the scope of the model. Two of these are shown in Figure 8. On the left, the data do not appear to be a "generalization" function at all; rather, they look much more like a probability density function or a typicality gradient, suggesting that this participant has interpreted the task in a manner more akin to a categorization problem (Ashby & Alfonso-Reese, 1995). That is, items that are clearly members of the concept, but likely to be on the fringes are in fact assigned low probability. In the right panel, the flat-topped region extends a long way to the right. Noting that the *data* are on the left side of the region, it would appear that this participants' prior is *not* location invariant. This is exactly the pattern one expects if one has a very strong prior bias for the region to be centered in the middle of the acceptable range.

Final Remarks

Even in very simple inductive tasks it is clear that people vary considerably in their prior beliefs and in their assumptions about how data are generated. When these effects are incorporated into Tenenbaum and Griffiths' (2001) generalization model, a number of counterintuitive effects can arise (e.g., concave curves). Nevertheless, we note that some characteristics of the model remain invariant (e.g., gradients may not become shallower with data), allowing quite stringent experimental tests of the theory. We present the first such test of the model, and show that it performs well in 36 of 66 cases, but cannot capture the full range of behaviors observed even in this simple task. While a number of post hoc extensions are possible, they are somewhat beyond the scope of this paper.

Acknowledgements. DJN was supported by an Australian Research Fellowship (ARC grant DP0773794). We thank Nancy Briggs for many helpful discussions, and Peter Hughes for writing the software used in the experiment.

References

Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology*, *39*, 216-233.

Baranski, J. V., & Petrusic, W. M. (1998). Probing the locus of confidence judgments: Experiments on the time to determine confidence. *Journal of Experimental Psychology: Human Perception and Performance*, *24*, 929-945.

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, *44*, 190-204.

Navarro, D. J. (2006). From natural kinds to complex categories. In *Proceedings of the 28th annual conference of the cognitive science society* (pp. 621–626). Mahwah, NJ: Lawrence Erlbaum.

Navarro, D. J. (2008). *Evaluating Bayesian theories of generalization.* (Available on the author's website)

Navarro, D. J., Griffiths, T. L., Steyvers, M., & Lee, M. D. (2006). Modeling individual differences using Dirichlet processes. *Journal of Mathematical Psychology*, *50*, 101–122.

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, *22*(4), 325-345.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*, 1317-1323.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*, 629-641.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327-352.

Weber, N., & Brewer, N. (2004). Confidence-accuracy calibration in absolute and relative face recognition judgments. *Journal of Experimental Psychology: Applied*, *10*, 156-172.